Capstone-1
Supplier Pricing Prediction and Segmentation

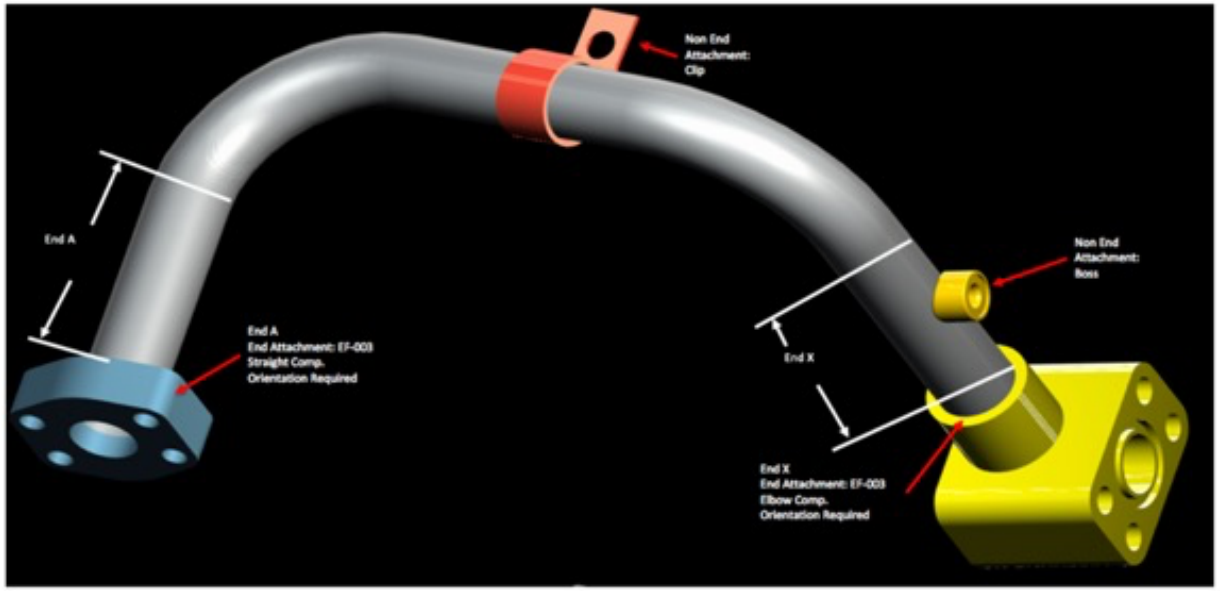
Domain Area: Supply Chain | Data Science

Motivation behind this Project?

Caterpillar (manufacturer of construction equipment) buys 8,855 unique tube assemblies of varying specifications from 57 different suppliers. Every time, company needs a tube assembly, they have to identify best qualified supplier and lowest best price, which has resulted in 30,213 requests for supplier quotes in 35 years.

Business Question:

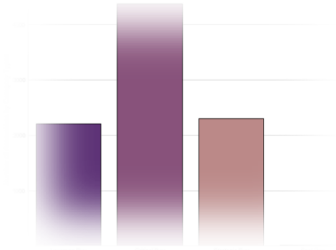
- Can we predict supplier pricing to save time and costs?
- Can we identify best managed assemblies and suppliers so that we can continuously improve?



About Assemblies: Which ones are Strategic?

4 Categories:

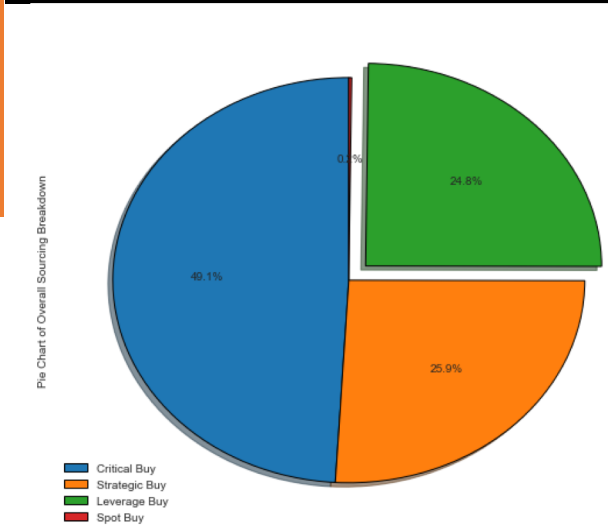
- Critical
- Strategic
- Leverage
- Spot

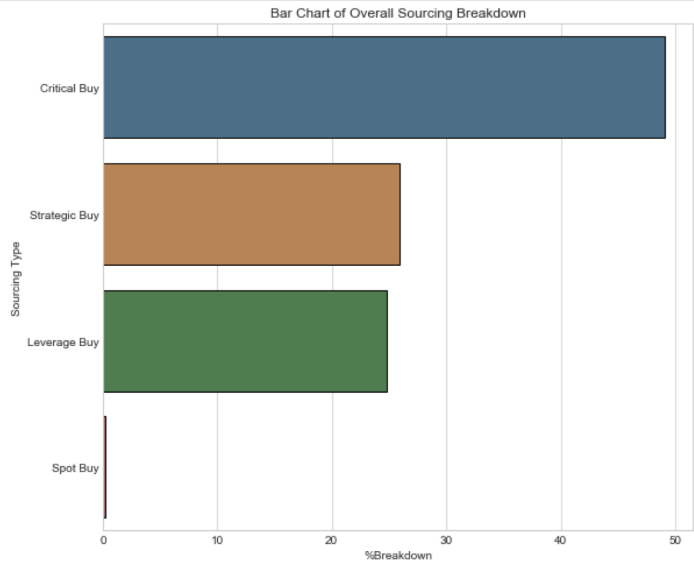


Years of High Spend : > 100 K

- 2005
- 2006
- 2009
- 2012
- 2013
- 2014

Assembly Category Breakdown:

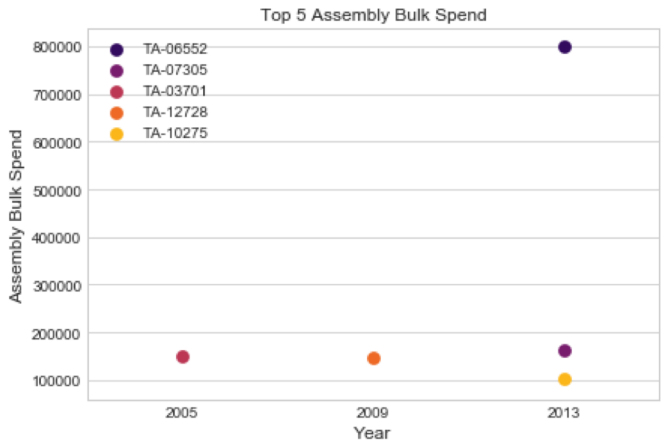


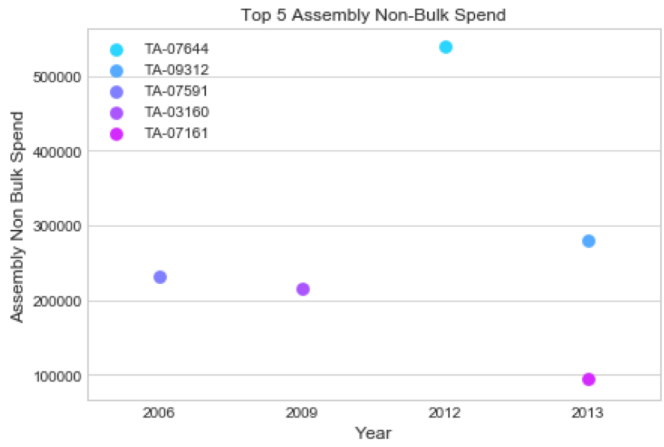


Observation:

1. Majority of tube assemblies are purchased from sole suppliers, 49.1%
2. Strategic and Leverage Options account 25.9% to 24.8% each.
3. There are opportunities to consolidate critical spend with leverage and strategic spends.

Top 5 Bulk and Non-Bulk Assembly Spends

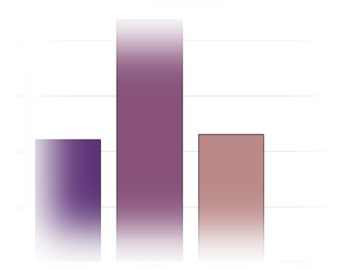




About Suppliers: Any Preferred Partners?

3Categories:

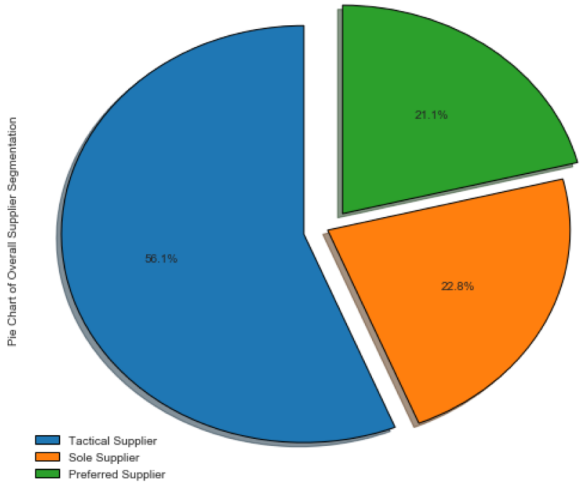
- Tactical
- Sole
- Preferred

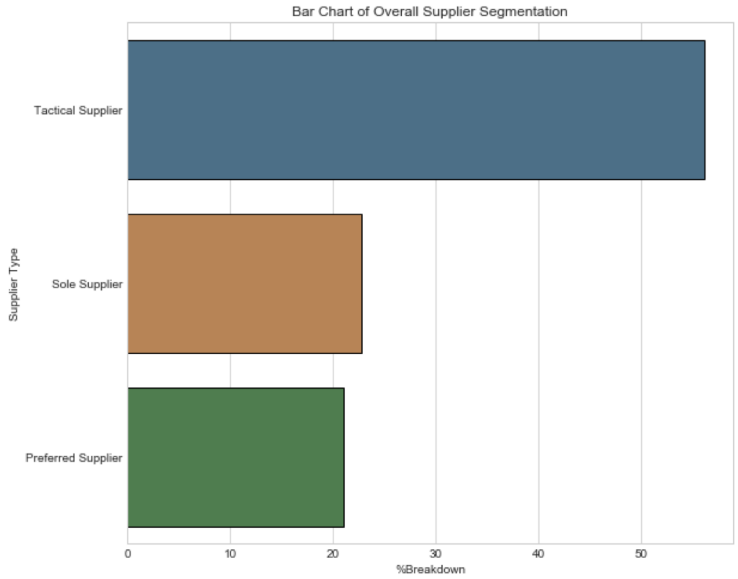


Pricing Types:

- Bulk Pricing
- Minimum Order Pricing

Supplier Segmentation:

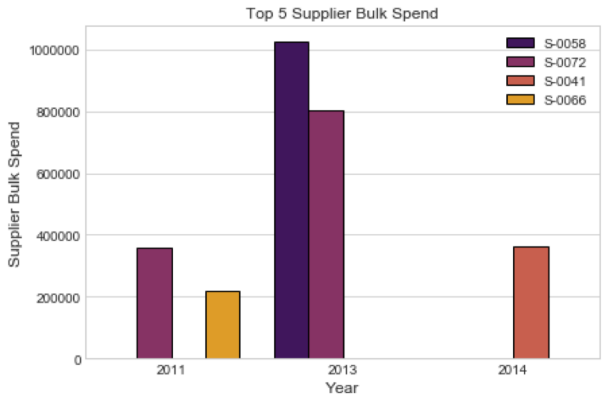


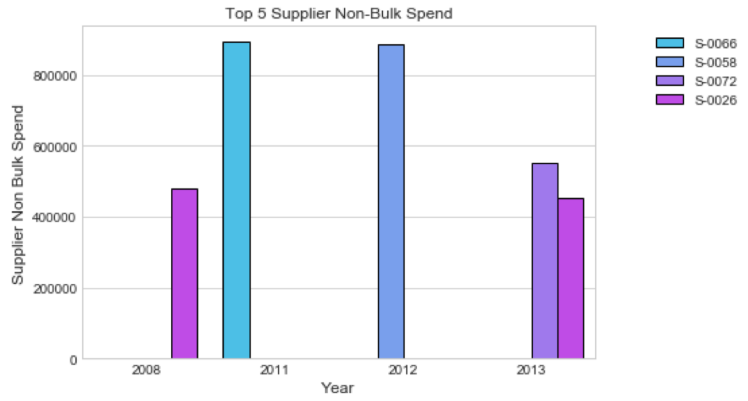


Observation:

1. High number of Tactical Suppliers, 56.1%
2. Fewer Preferred suppliers than Sole Suppliers.
3. There may be opportunities to transform tactical suppliers into Preferred.
4. May be an opportunity to conduct business review why we sole source?.

Top 5 Bulk and Non-Bulk Supplier Spends

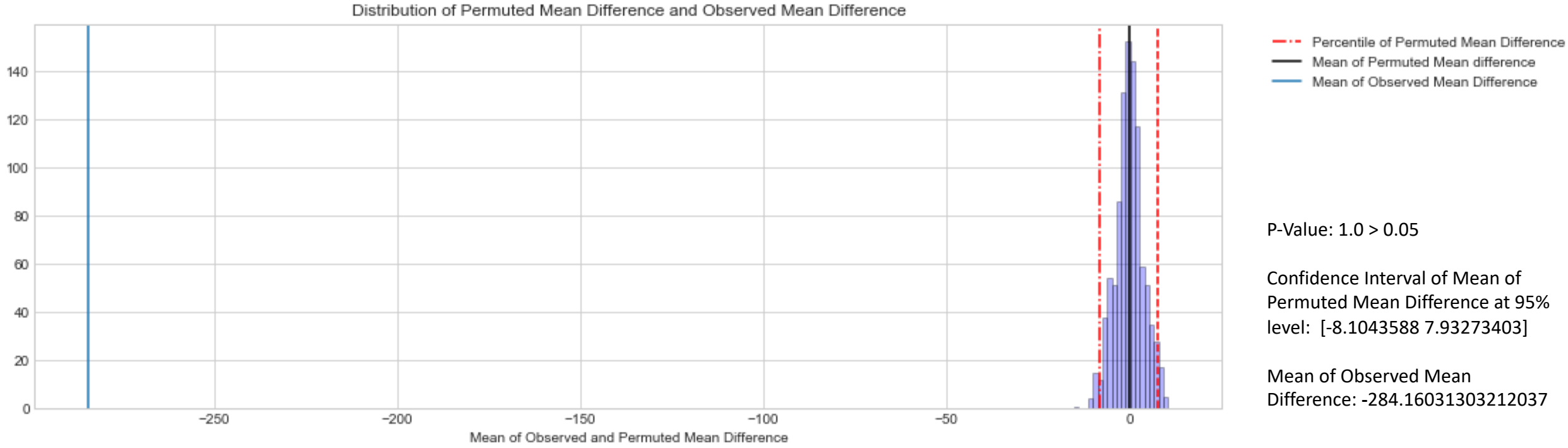




Is Buying Bulk Assemblies on Minimum Order Basis a significant change?

Tested Hypothesis by simulating 1000 samples as shown in the graph below.

- Null Hypothesis: Business will continue buying bulk assemblies in bulk.
- Alternate Hypothesis: Business will buy bulk assemblies as non-bulk (i.e. on minimum order basis)
- Calculate mean difference between observed and permuted samples for bulk and non-bulk assemblies.



P-Value: 1.0 > 0.05

Confidence Interval of Mean of Permuted Mean Difference at 95% level: [-8.1043588 7.93273403]

Mean of Observed Mean Difference: -284.16031303212037

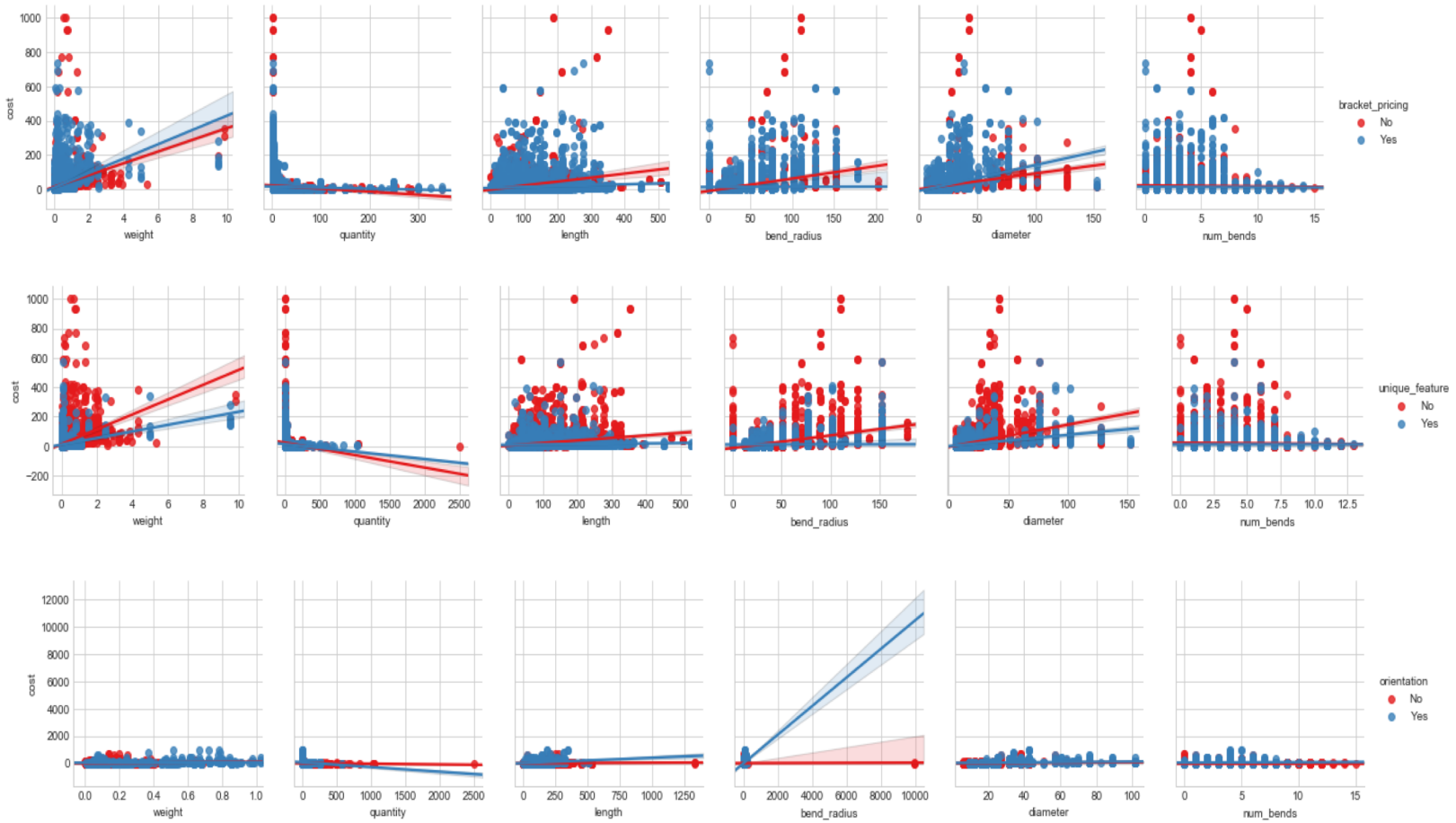
Observation:

1.High p-value value means that null hypothesis exists, even if we took 1000 samples to run the simulation.
2.Bulk assembly parameters will be different from the non-bulk assembly parameters, unless there is a significant shift in the observed mean difference.
3.Significant shift in the observed mean would mean that the bulk assembly parameters such as weight, costs and quantity have changed resulting in bulk assemblies being sourced as non-bulk (minimum order basis), which can be tracked and compared with the collected data.
4.When mean of observed mean difference becomes > 7.93, it lowers the p-value under 0.05, showing discontinuation in the business trend.

Input Parameters for Predicting Supplier Pricing

- Weight
- Length
- Diameter
- Bend Radius
- Number of Bends
- Orientation
- Pricing Type
- Unique Features
- Added Features
- Suppliers
- Components
- Material Specs
- Tube End Type
- Formed/Unformed Ends
- Quantity
- Usage
- Pricing Type

Dependency of Cost on Few Input Features

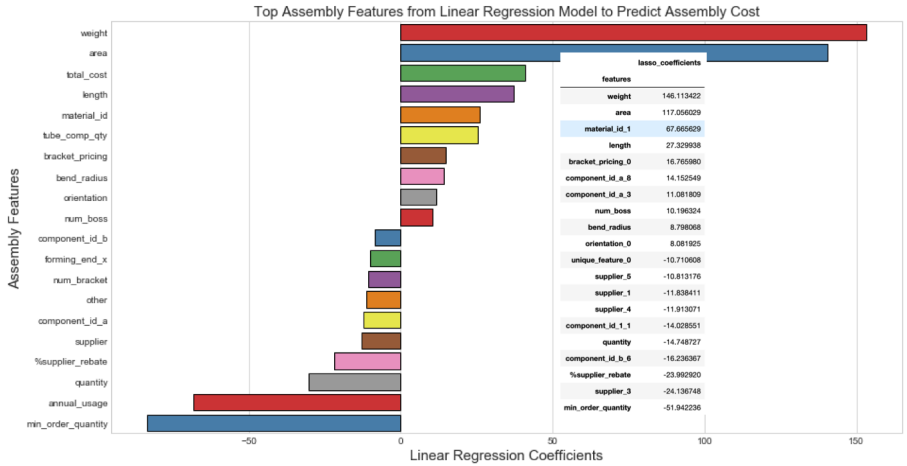


Observation:

1. Cost increases with increase in weight/length/bend radius/diameter and decreases with the increase in order quantity.
2. Cost decreases with increase in order quantity.
3. Heavier and bigger diameter bulk assemblies have higher cost increase.
4. Cost is not effected by the increase in number of bends.
5.. Assemblies with Unique features have lower cost increase. Could be due to lower demand.
6. Cost of having unique features is not impacted by increase in length, bend radius and number of bends.
7. Bulk assemblies with orientation have steep cost decrease.
8. Assemblies which are oriented and have Longer length and higher bend radius are more costly.
9. Orientation has negligible impact on assemblies weight and tube diameter.

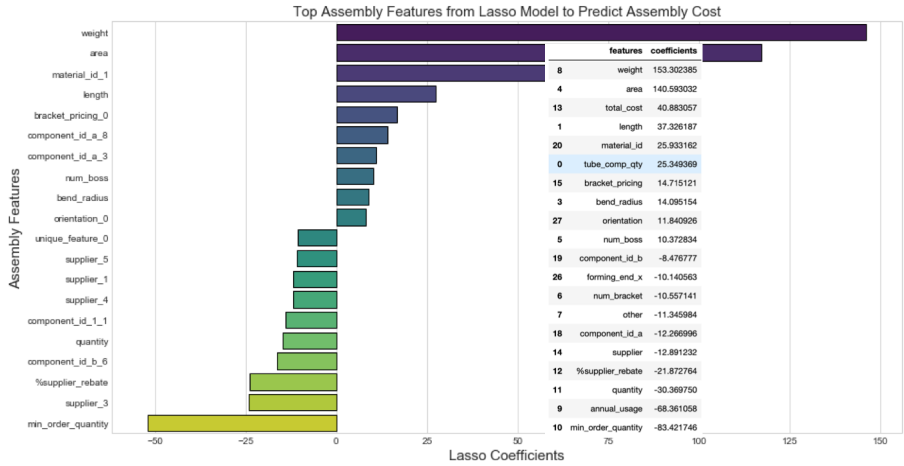
Feature Importance Comparison Linear Model Vs Random Forest

Linear Regression: Top 20 features, Explained Variance 32%



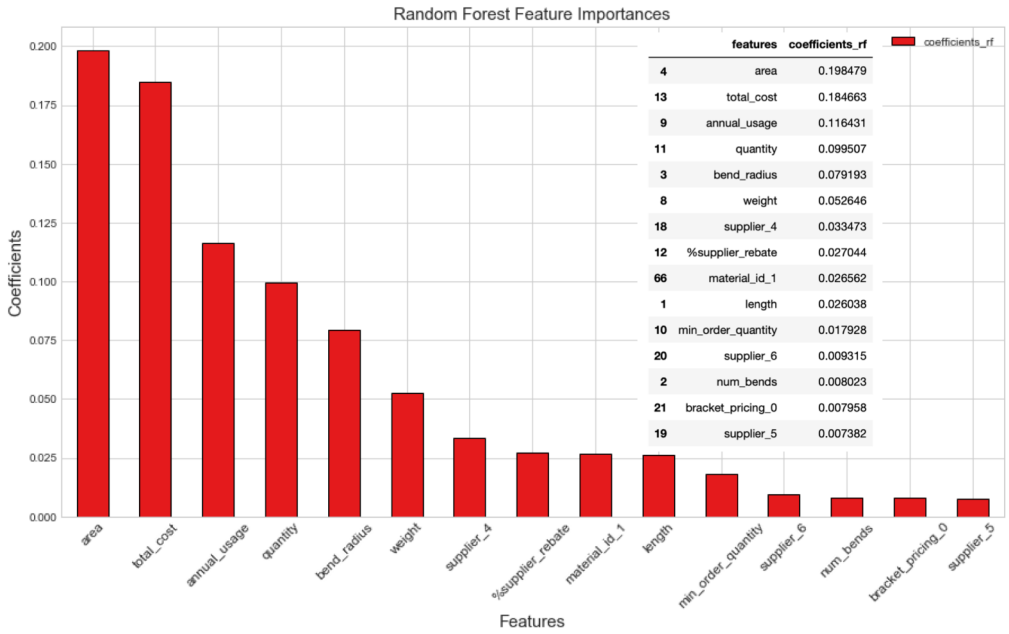
features	lasso_coefficients
weight	146.113422
area	117.056029
material_id_1	67.665629
length	27.329938
bracket_pricing_0	16.765980
component_id_a_8	14.152549
component_id_a_3	11.081809
num_boss	10.196324
bend_radius	8.798068
orientation_0	8.081925
unique_feature_0	-10.710608
supplier_5	-10.813176
supplier_1	-11.838411
supplier_4	-11.913071
component_id_1_1	-14.028551
quantity	-14.748727
component_id_b_6	-16.236367
%supplier_rebate	-23.992920
supplier_3	-24.136748
min_order_quantity	-51.942236

Lasso: Top 20 features, Explained Variance 42%



	features	coefficients
8	weight	153.302385
4	area	140.593032
13	total_cost	40.883057
1	length	37.326187
20	material_id	25.933162
0	tube_comp_qty	25.349369
15	bracket_pricing	14.715121
3	bend_radius	14.095154
27	orientation	11.840926
5	num_boss	10.372834
19	component_id_b	-8.476777
26	forming_end_x	-10.140563
6	num_bracket	-10.557141
7	other	-11.345984
18	component_id_a	-12.266996
14	supplier	-12.891232
12	%supplier_rebate	-21.872764
11	quantity	-30.369750
9	annual_usage	-68.361058
10	min_order_quantity	-83.421746

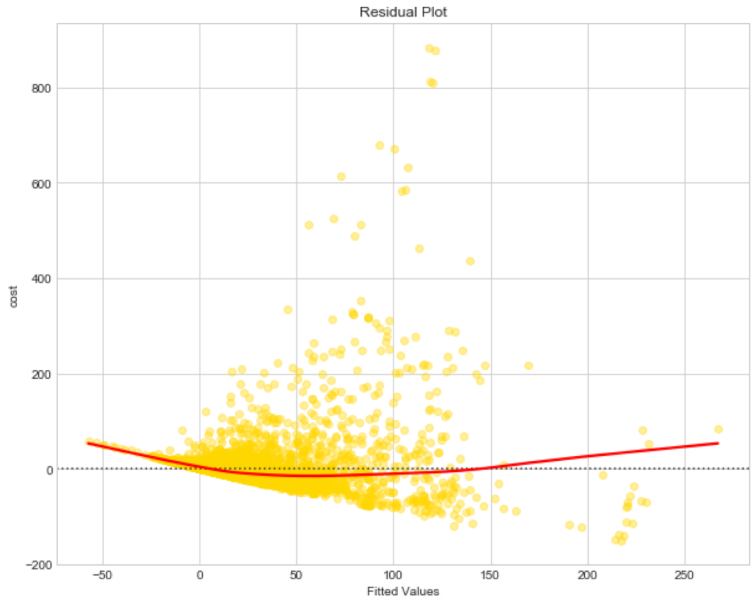
Random Forest: Top 15 features, Explained Variance 89%



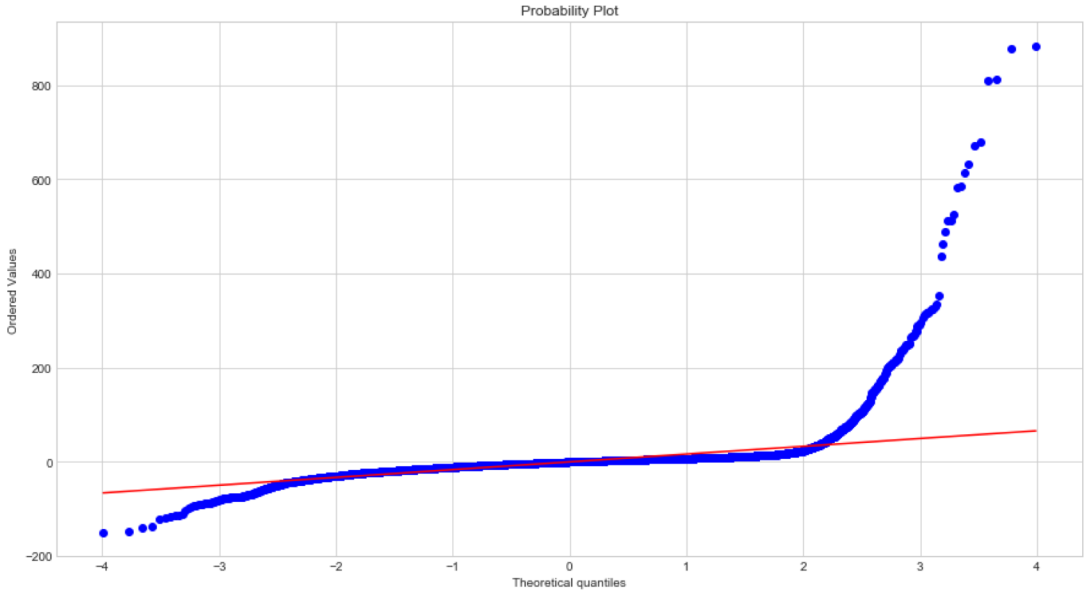
	features	coefficients_rf
4	area	0.198479
13	total_cost	0.184663
9	annual_usage	0.116431
11	quantity	0.099507
3	bend_radius	0.079193
8	weight	0.052646
18	supplier_4	0.033473
12	%supplier_rebate	0.027044
66	material_id_1	0.026562
1	length	0.026038
10	min_order_quantity	0.017928
20	supplier_6	0.009315
2	num_bends	0.008023
21	bracket_pricing_0	0.007958
19	supplier_5	0.007382

Reason for lower explained variance in Linear Models

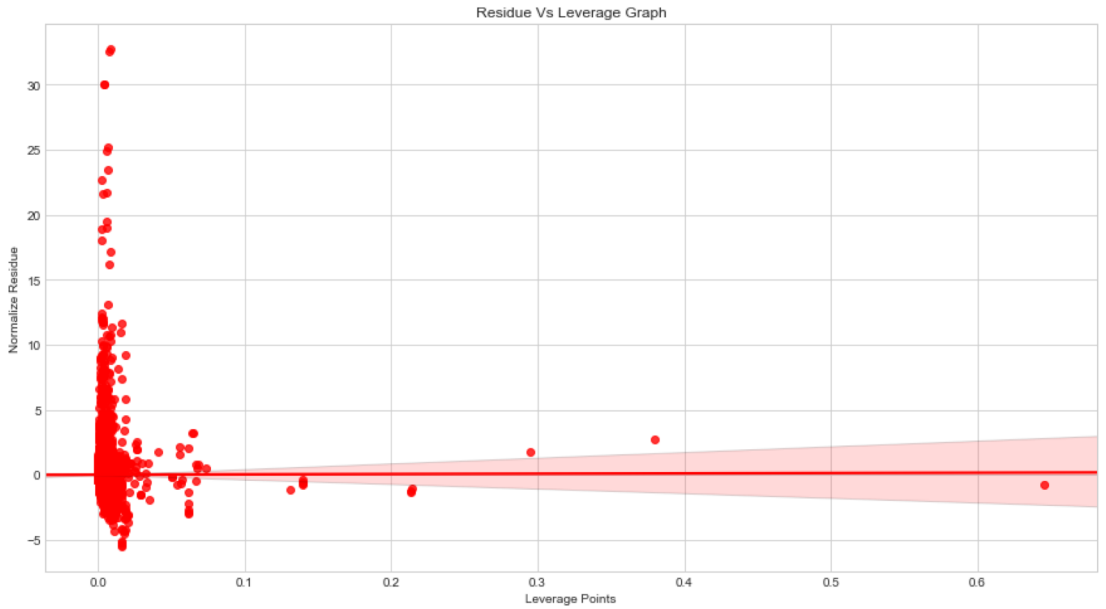
Skewness in Residual Distribution Plot



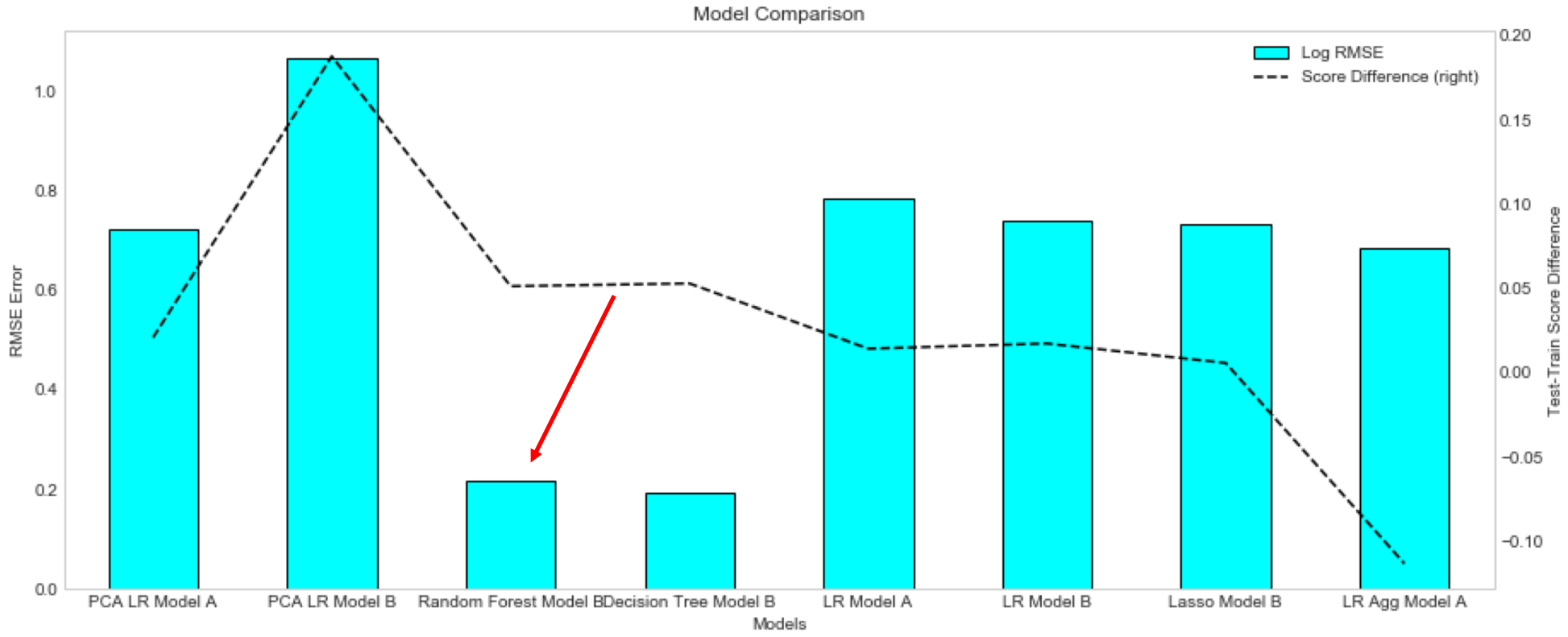
Outliers in Residual Values



Leverage Points in Residual Values



Supplier Price Prediction: Model Evaluation



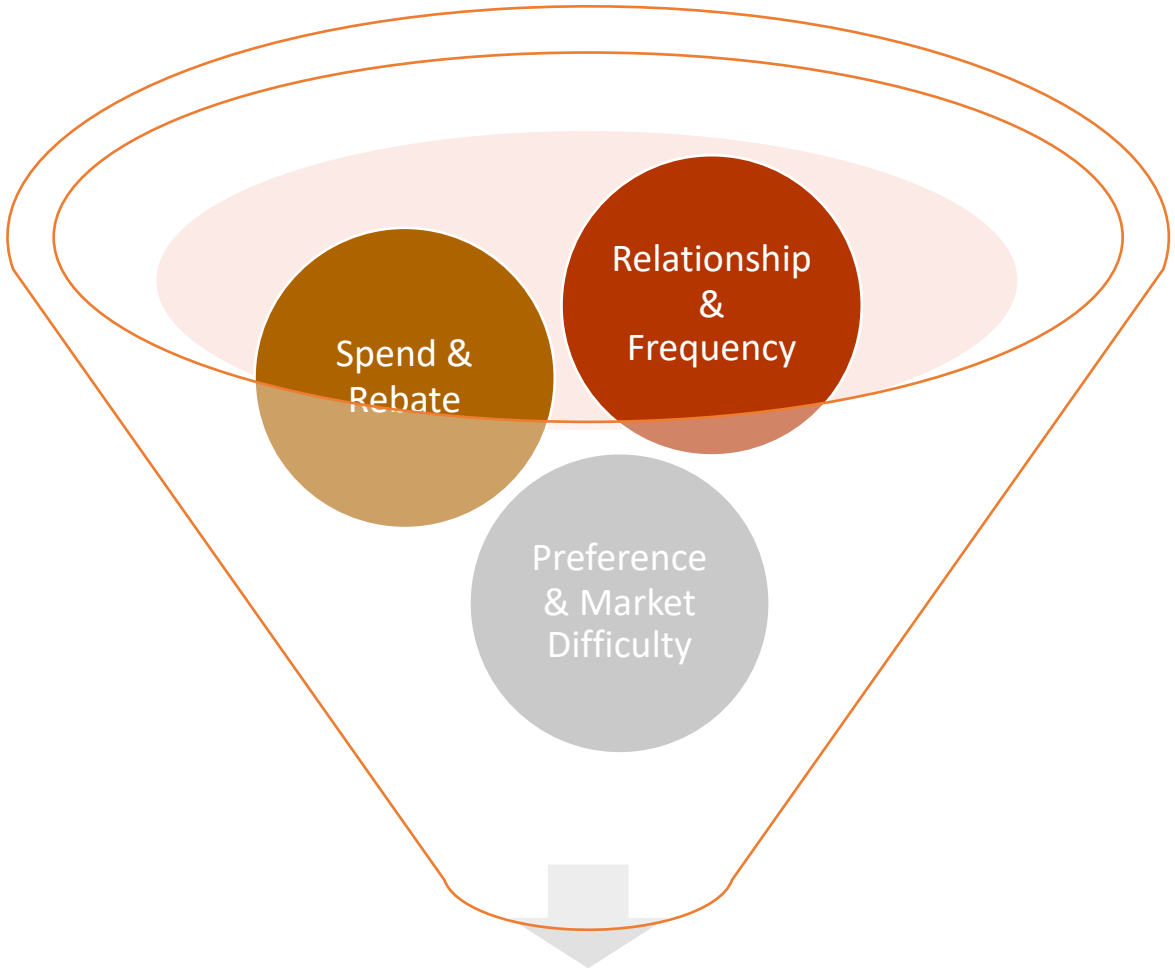
Observation:

1. Random Forest resulted in low log RMSE value of 0.21 and Train-Test score difference of 5.1%.
2. Linear models had comparable log RMSE values except PCA LR on model B because of categorical transformation. .
3. **Random Forest, after hyperparameter tuning resulted in 98% accuracy and gave top 15 features, which explained 89% variance in the predicted cost.**

Criteria for Segmenting Assemblies and Suppliers

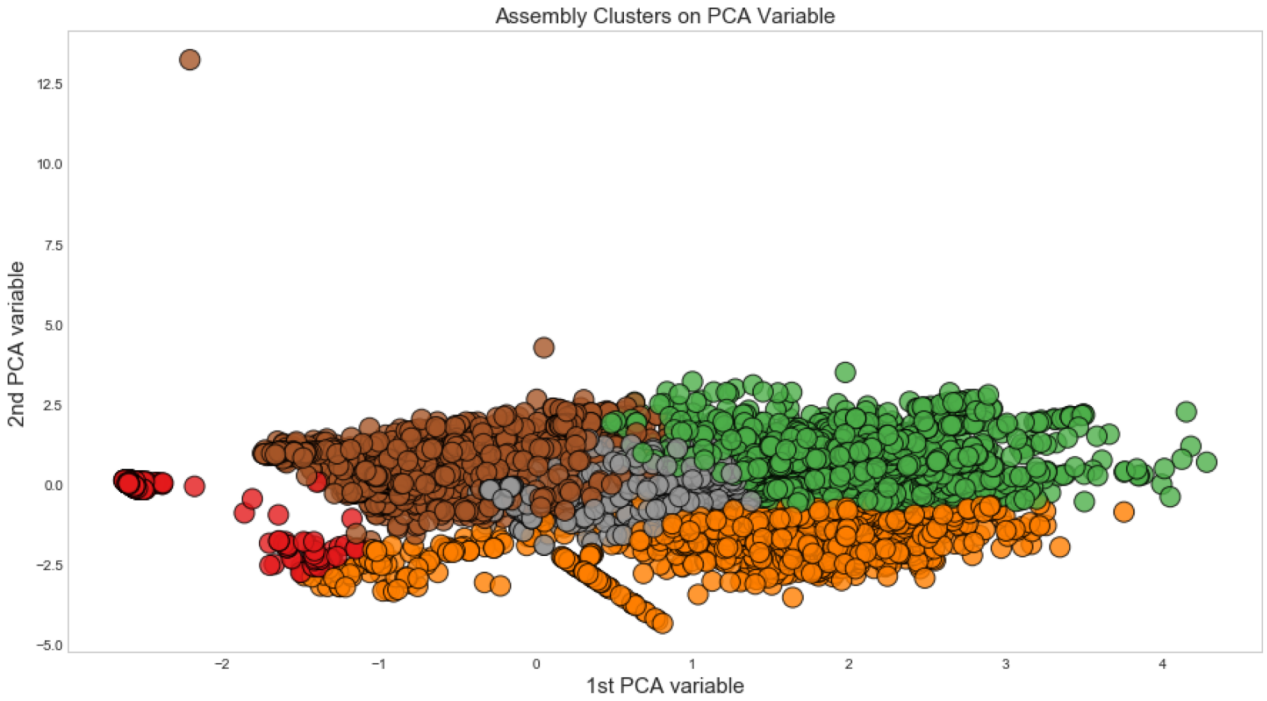
Business Needs:

- Spend: Total Spend Bulk/Non-Bulk
- Rebate: Discounts
- Relationship: Length of Contract
- Frequency: Number of Times Purchased
- Market Difficulty: Ease in switching supplier
- Preference: Supplier with maximum assortment
- Recency: Most Recent Purchase

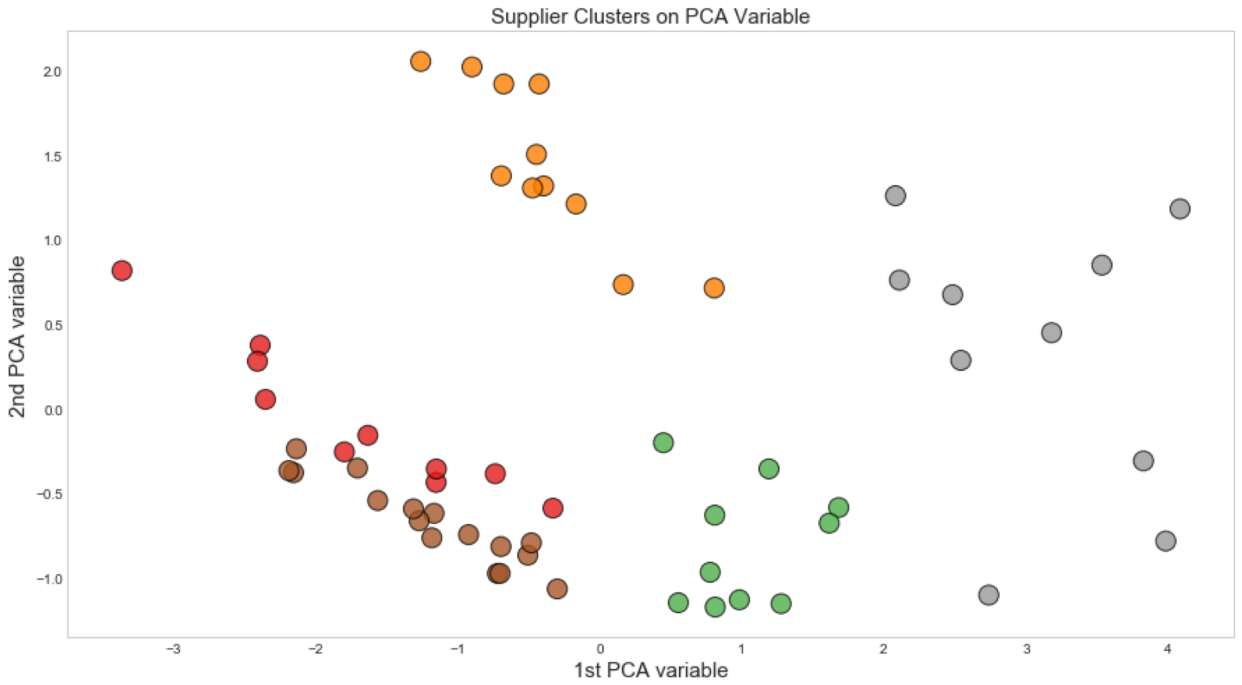


Cluster Quality

Assembly Clusters = 5
Silhouette Score 0.46
Model: K-Means
Good Dense Formation

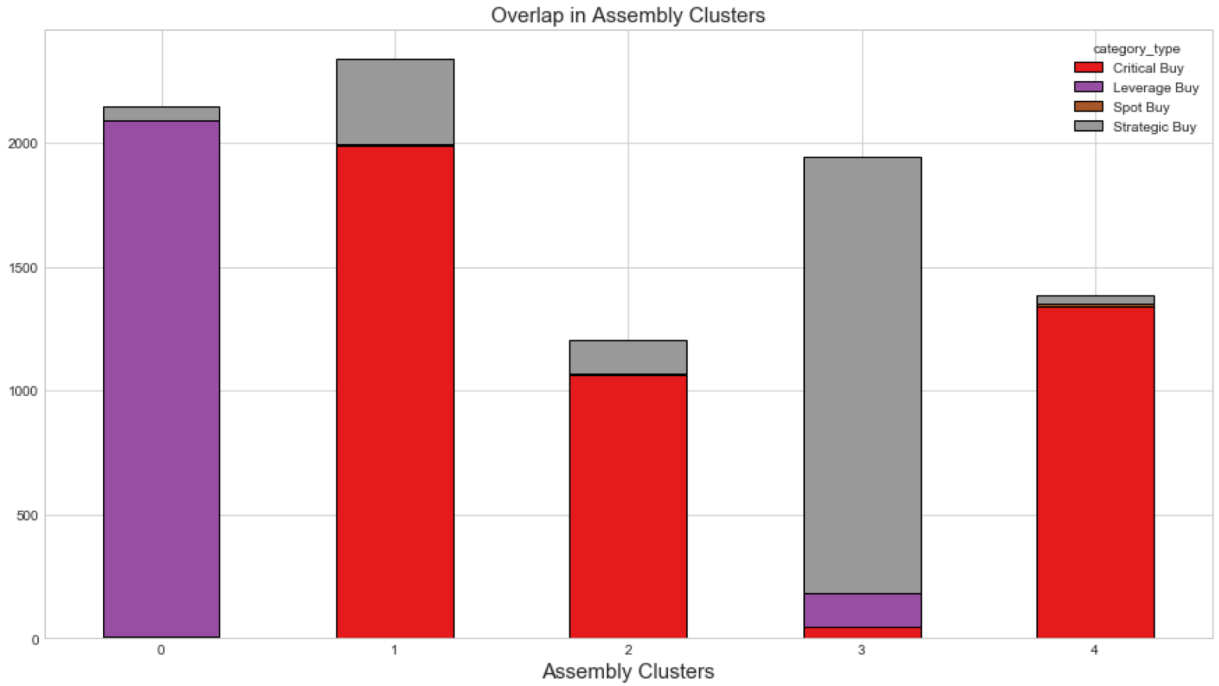


Supplier Clusters = 5
Silhouette Score 0.35
Model: K-Means
Less Dense Formation

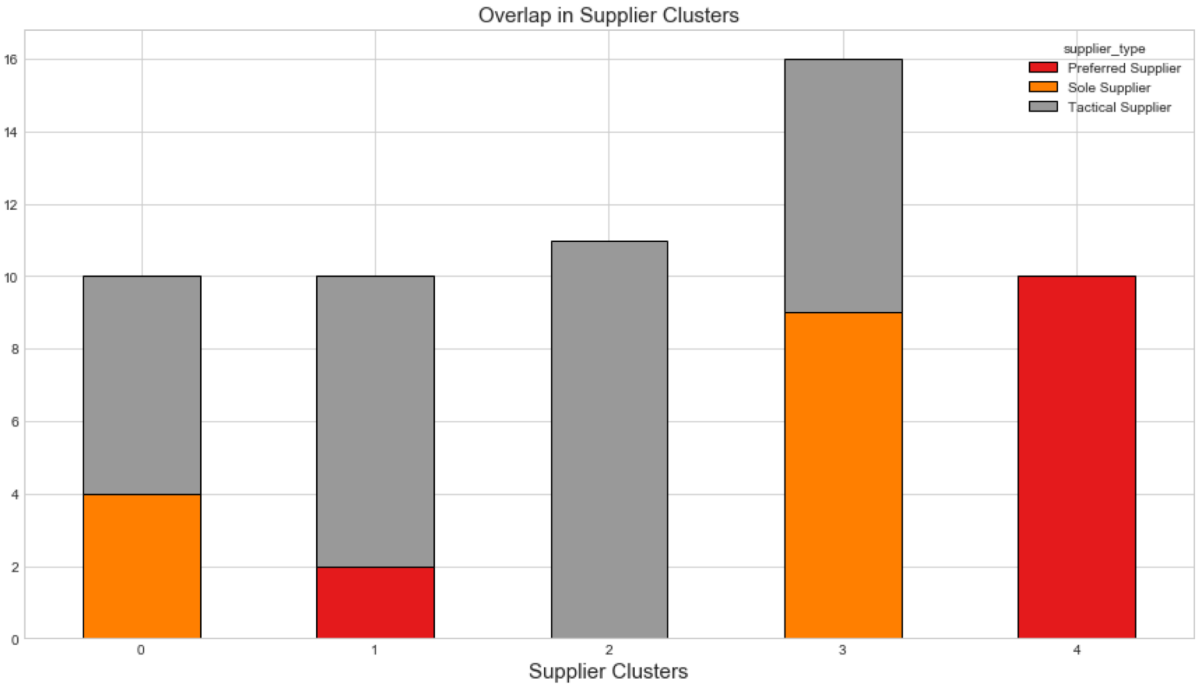


Cluster Overlap

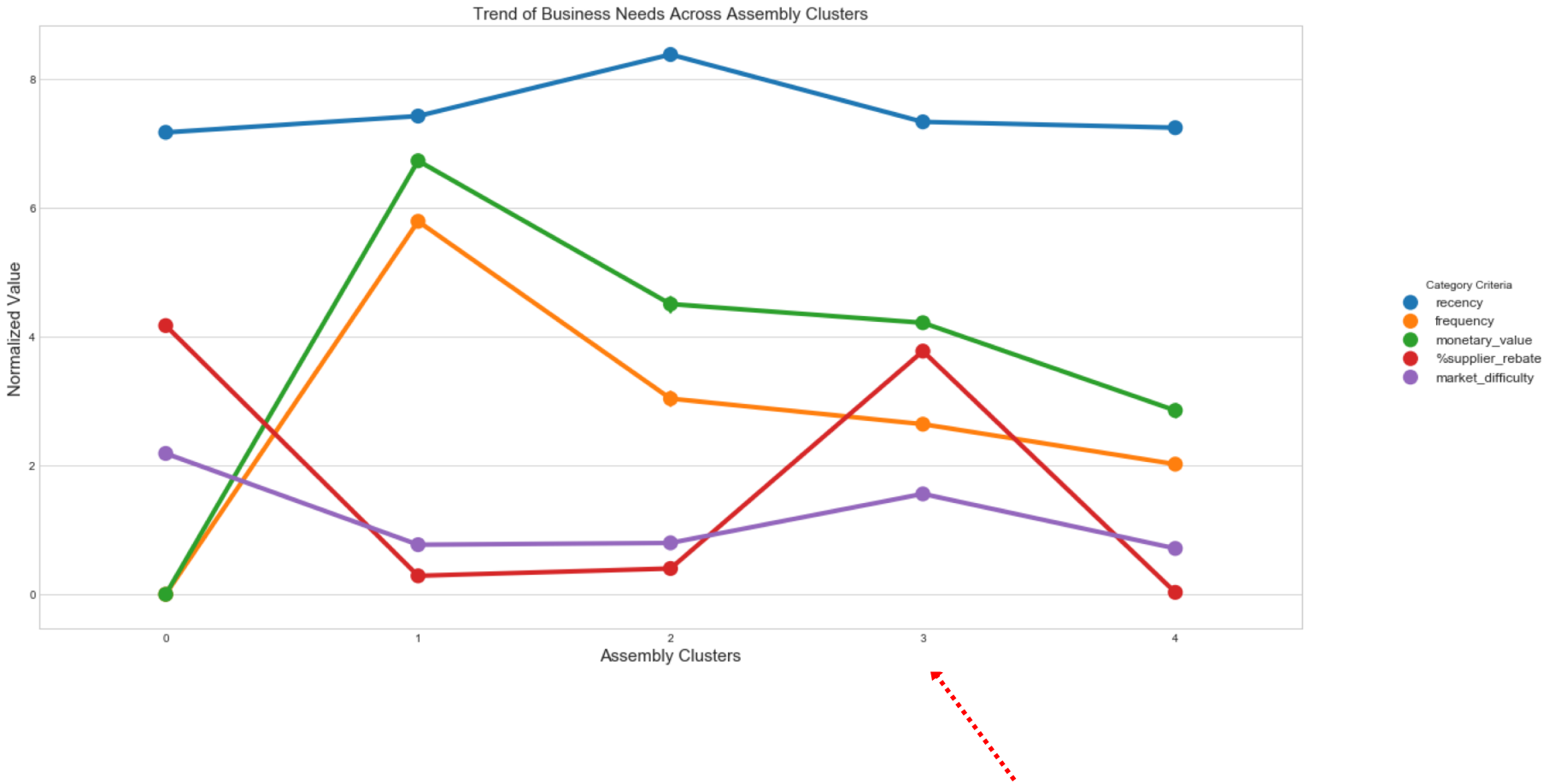
All Assembly Categories have some Strategic Buy



Some Tactical suppliers are Sole as well as Preferred



Best Managed Assembly Cluster #3

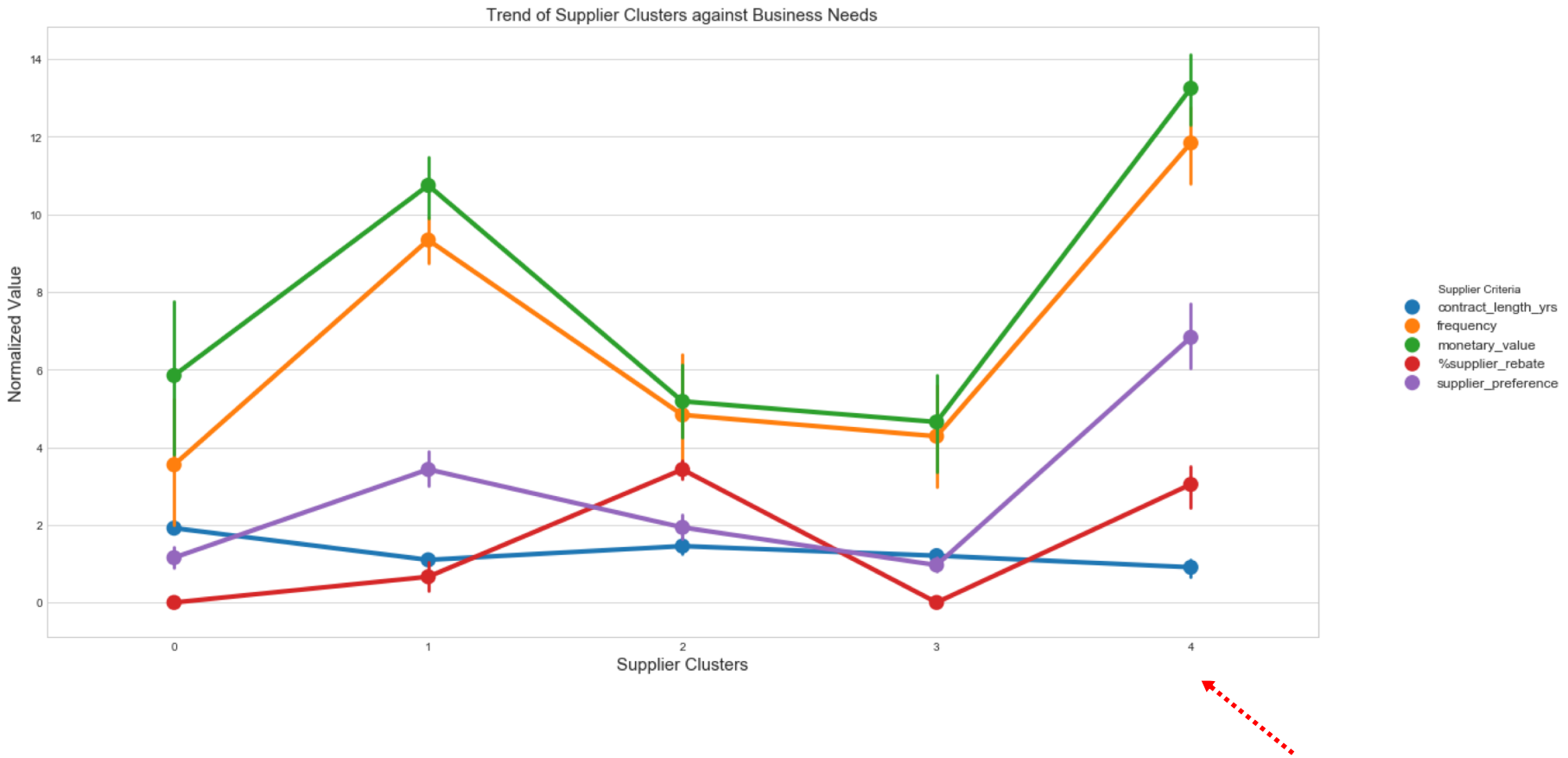


Observations:

1. Assembly Clusters 1,2 and 4 have high supplier spend and frequency of purchase but lowest supplier rebates.
2. Assembly Cluster 0 does not have supplier spend but are motivated to offer high rebates (potential for leverage buy).
3. Assembly Cluster 3 is the best managed category offering most rebates have fewer suppliers (like strategic buy).

Assembly Cluster 3 can be used as a benchmark to manage other clusters and find improvements.

Best Managed Supplier Cluster #4



Observations:

1. Supplier clusters 0 and 3 has the lowest supplier preference and offer no rebates (similar to Sole Suppliers). We should consider possible scope re-assignment or spend consolidation with tactical or preferred suppliers in order to obtain better rebates.
2. Supplier Cluster 4 is the best managed portfolio offering high rebates and have high supplier preference (Similar to Preferred Suppliers).
3. Supplier Cluster 2 demonstrates similar characteristics with cluster 4 but demonstrates very low supplier preference. There may be an opportunity to develop cluster 2 at par with cluster 4 and encourage healthy competition.
4. Supplier cluster 1 has shorter contract length, 2nd best supplier preference and significantly lower supplier rebates relative to high spend they manage. There should be an opportunity to negotiate better contract rates for a longer term deal.

Supplier Cluster 4 can be used as a benchmark to manage other suppliers and find improvements.

Project Conclusion

- Using Random Forest and K-Means, we were able to establish pricing prediction and benchmark assemblies and supplier clusters using 8,855 unique assemblies and 57 suppliers.

- There were top 15 features such as tube assembly area, total cost, annual usage, order quantity and more, which explained 89% dependency in predicting supplier pricing.

- By combining prediction and clustering analysis business can benefit by estimating supplier pricing based on assembly specifications, demand and order history; identify assembly and supplier portfolios by business needs; and choose suppliers that meet business requirements.

Future Possibilities

- Monitor Supplier Performance: We can add other features such as on-time delivery, safety statistics, contract compliance and inventory levels to gain further supplier insights.

- Build Real Time Dashboards: We can connect our model to real time data feed to collect actionable insights on the fly.

- Further Modeling: We can try other algorithms such as time-series to predict annual supplier pricing.